第 37 卷第 3 期
录用定稿
计算机应用研究
Application Research of Computers
Vol. 37 No. 3
Accepted Paper

融合模拟退火的随机森林房价评估算法 *

丁旸钧天，曹怀虎

(中央财经大学 信息学院，北京 100081)

摘　要：传统的随机森林房价评估算法存在着大量参数组合计算问题，参数的优劣对算法准确度影响很大。针对此问题，结合随机森林和模拟退火算法，提出一种融合模拟退火的随机森林房价评估算法。首先，通过 10 次十折交叉验证法对参数进行敏感性测试，选择出对随机森林算法敏感的参数；然后，结合模拟退火算法对敏感的参数迭代寻优，通过与网格搜索算法、随机搜索算法进行对比分析，发现在参数组合计算过程中，模拟退火算法在运行时间和算法准确率方面优于网格搜索算法与随机搜索算法，弥补了网格搜索算法耗时过长和随机搜索算法低准确率的缺陷；最后，将融合模拟退火的随机森林算法应用于房价评估问题，构成新的房价评估算法。将新算法与传统随机森林房价评估算法进行了对比实验分析，结果表明，融合模拟退火的随机森林房价评估算法误差值减少，拟合优度值增加，评估的准确度得到了显著提升。

关键词：随机森林；模拟退火；参数优化；房价评估

中图分类号：TP301.6　　doi: 10.3969/j.issn.1001-3695.2018.07.0613

Housing prices evaluation using random forest algorithm combing with simulated annealing

Ding Yangjuntian, Cao Huaihu

(*School of Information*, *Central University Finance & Economics*, *Beijing* 100081, *China*)

Abstract: The traditional housing prices evaluation which was using Random Forest algorithm had a large number of parameter selection problems. The parameters had great influence on the accuracy of the algorithm. In order to solve this problem, this paper combined the Random Forest algorithm and simulated annealing algorithm and proposed a new algorithm about the housing prices evaluation. Firstly, according to the different sensitivity of the Random Forest parameters to the algorithm, this paper tested the sensitivity of the parameters by 10 times 10-cross-validation method, then selected the parameters of the algorithm. Secondly, this paper used the simulated annealing algorithm to the sensitive parameters iterative optimization. Through comparing to the grid search algorithm and random search algorithm, this paper found the simulated annealing algorithm do better than the grid search algorithm and random search algorithm in the running time and algorithm accuracy. The simulated annealing algorithm made up the defects of the time-consuming and the low-accuracy of the random search algorithm in the grid search when selecting parameters. At last, this paper applied the Random Forest algorithm combing with simulated annealing to the problem of housing prices evaluation, and formed a new evaluation algorithm. Comparing the new algorithm with the traditional Random Forest price estimation algorithm, the results show that the error value of the Random Forest price estimation algorithm with simulated annealing is reduced, the goodness of fit value increases, and the accuracy of the evaluation is improved markedly.

Key words: random forest; simulated annealing; parameter optimization; housing prices evaluation

0　引言

随着经济发展与城镇化进程的不断推进，越来越多的人将房产视为一种投资，房地产行业日渐火爆，房产交易量日益增加。作为房产交易的必然环节，房价评估受到广泛重视。传统房价评估方法如市场比较法、成本法、交易法和回归算法预测法。在小数据量的房价评估中，传统评估方法准确度较高。随着数据规模增长，传统方法需要大量的计算成本与人工成本。随着机器学习的发展，针对房价评估问题，文献[1]中首次提出应用神经网络技术进行房价评估。文献[2]提出应用支持向量机进行房产评估，发现支持向量机算法可以获得良好的预测效果。文献[3]中首次提出应用随机森林的方法进行房价评估，发现随机森林可以提高预测的准确度。

然而以上研究中忽略了参数选择对于算法的影响。文献[4]表明不同预测算法所达到的最佳性能的参数设置不同，参数调优是算法优化的重要一步，对于随机森林算法，参数调优同样重要。常用的参数调优方法为网格搜索法[5]和随机搜索法[6]。网格搜索法类似于穷举，准确度较高，但在参数范围较大的数据中需要耗费大量的时间，大大降低了算法性能；随机搜索法通过随机抽样寻找最优解，在时间效率方面要远远优于网格搜索法，但该方法随机性太强，容易陷入局部最优解。文献[7]通过优化决策树的数量选择更高准确度的子树，提高算法预测准确度。文献[8]通过对 OOB 误差最小化处理进行超参数估计，拟获得近似最优解。文献[9]中改进的网格搜索进行参数优化，保证了搜索到近似最优组合周边所有可能区域，提高了网格搜索的时效性，但对于更大数据量的问题时效性仍不高。

为提高随机森林房价评估算法参数寻优的效度与评估准

收稿日期：2018-07-31；**修回日期：**2018-10-15　**基金项目：**北京市社会科学重点项目（16YJA001）；国家自然科学基金项目（61671030）

作者简介：丁旸钧天（1994-），女，河北张家口人，硕士研究生，主要研究方向为机器学习，金融科技（ellie_dy@sina.com）；曹怀虎（1977-），男，教授，博士，主要研究方向为网络经济学、社会计算.

确度，本文提出融合模拟退火的随机森林算法，利用模拟退火逐步降温，迭代寻优的特点，将算法融合到传统的随机森林房价评估算法中，进行参数寻优与特征选择。首先，根据随机森林参数对算法敏感性不同，运用十折交叉验证法对参数进行敏感性测试，选择出对算法敏感的参数；其次，通过模拟退火算法对敏感的参数迭代寻优，并与网格搜索算法、随机搜索算法进行对比分析，发现，在参数组合计算过程中，模拟退火算法在运行时间和算法准确率方面优于网格搜索算法与随机搜索算法，弥补了网格搜索算法高耗能和随机搜索算法低准确率的问题；最后，将融合模拟退火的随机森林算法应用于房价评估问题，构成新的房价评估算法。将新算法与传统随机森林房价评估算法进行了对比实验分析，结果表明，融合模拟退火的随机森林房价评估算法误差值减少，拟合优度值增加，评估算法准确度得到了提升。

1 相关研究

1.1 房价评估传统方法

房价评估传统方法有市场比较法、成本法和交易法，以及通过回归算法进行房价预测的方法。市场比较法[10]是常用的经验估值法，通过与周边房价进行对比，得出符合自身房屋价值的价格。市场比较法是房价评估常用方法，但是这种方法需要大量数据以及丰富的经验，且前提假设为房价稳定且没有市场垄断的情况，实际生活中房价处于波动状态，且难以避免有垄断情况的出现。成本法[11]即建造成本加上各项税费和正常的利润进行房价评估的方法。成本法适用于房地产交易较少的情况，而且成本法中的房地产开发商的利润往往与实际情况有所出入。交易法[12]指预计估价对象未来各期的正常净收益，选用适当的资本化率将其折算到估价时点上的现值后累加，以估算估价对象的客观合理价格或价值的方法。交易法适用于评估有收益或潜在收益的房地产，但其中的折旧率的计算拥有很大争议。除了传统房价评估三大方法，也有不少学者应用回归算法对房价进行相关分析，如文献[13]中利用面板分位数回归算法进行实证分析，表明收入是拉动我国大中城市房价的主要因素。文献[14]中运用地理加权回归(GWR) 算法，探索土地供应政策对房价的影响机制。研究表明，其他条件不变的前提下，土地供应量和结构对房价具有显著的负效应。

1.2 传统的随机森林房价评估算法

随机森林回归算法[15]是 Bagging 算法[16]的发展，是多个弱学习器输出为强学习器的过程。传统的随机森林房价评估算法，主要通过用户输入相关信息，借助随机森林回归算法对房价进行评估。主要包括如下几步：

a) 数据采集。通过网络爬虫或者现有软件对数据进行采集。

b) 数据预处理。对数据进行整理形成数据集 $D=\{(x_1,y_1),(x_2,y_2),\ldots(x_m,y_m)\}$ 。

c) 特征选择。运用相关算法如决策树、逻辑回归等选择出对算法影响较大的特征变量，并更新数据集。

d) 算法训练。将数据集与初始化参数进行算法训练，形成强学习器算法。

e) 算法应用。通过用户输入信息，对其房价进行评估。

1.3 模拟退火算法

常用的参数调优算法有网格搜索法、随机搜索法。网格搜索算法类似于穷举算法，将所有参数结果进行组合计算，选出最优值。随机搜索算法增加了随机的特性，选取部分参数值进行最优选择，然而这种算法容易陷入局部最优解。参数调优过程类似寻找最优路径的过程，不少学者对最优路径的方法进行研究[17, 18]，基于前辈对于最优路径研究的启发，本文采用模拟退火算法解决算法参数寻优问题。

模拟退火算法(simulated annealing，SA)[19]模仿固体降温并寻找最优点的过程。初始温度 T 为最高温度，此时固体震荡最大；随着温度的降低，固体逐渐找到最优点并趋于稳定。从算法角度而言，模拟退火算法从初始温度 T 开始通过判断评价函数接受优于原函数的解或以波尔兹曼(Boltzmann)概率函数接受稍差一些的解，并逐步降温到 Tmin。SA 算法是一种启发式的搜索算法[20]，在爬山算法的基础上添加了概率函数，可以收敛到全局最优解，弥补了爬山算法陷入局部最优解的缺陷。模拟退火算法描述如下:

a) 初始化值 x。

b) 降低温度 T，计算新的解 x'，计算评价函数值。如果相对误差 $\Delta y<0$，或以波尔兹曼概率 $\exp(\Delta y/T)$ 接受新解，若接受，则令 $x=x'$。

c) 判断温度是否降到 Tmin 以下，或连续大量次数计算后没有更优解，则结束算法。

2 融合模拟退火的随机森林房价评估算法

2.1 敏感性参数生成

房价评估问题是回归问题的一种，本文选取随机森林算法解决此类问题，相对于传统算法，如线性回归与支持向量机，随机森林受参数的影响较大，参数的选择尤为重要，为了节省工作实效，提高算法运行效率，本文对敏感性参数进行提取，将敏感性参数加入算法的调优。随机森林算法的主要参数及其对算法的影响如下：

a) n_estimators。数据类型为 integer，表示随机森林中决策树的数量，文献[21]表明较多的决策树可以提高算法性能，但同时过多的决策树数量也会影响算法运行效率进而影响性能，到达一定数量后，性能趋于稳定。

b) max_features。数据类型为 int、float、string，表示训练集最大特征数，此参数增加一般能提高算法的性能，但是也降低了算法的速度以及单棵树的多样性，文献[22]表明当所有特征都进行分裂，反而会影响算法准确性。

c) min_samples_split。数据类型为 int、float，表示划分节点所需最小样本数，如果达到该值则不再进行划分，文献[23]表明此值对算法效果影响不大。

d) min_samples_leaf。数据类型为 int、float，表示叶子节点最小样本数，叶作为决策树的末端节点，较小的叶子更易于算法的降噪。

e) max_leaf_nodes。数据类型为 int，表示限制最大叶子节点数，可以防止过拟合。

f) max_depth。数据类型为 int，表示决策树的最大深度，取决于数据的分布情况，当大于此值则不再分裂。

针对不同问题，随机森林算法达到最优性能参数选择不同，各参数对算法敏感程度也不同，文献[1]采用交叉验证的方法进行参数敏感性测试，依此，本文采取 10 次 10 折交叉验证平均误差率对算法主要参数进行敏感性测试，挑选出针对房价评估问题的有效参数。

10 折交叉验证的方法是用来测试算法准确性的常用方法。将数据集分成 10 份，轮流将其中 9 份作为训练数据、1 份作为测试数据进行实验。每次实验都会得出相应的正确率（或差错率）。10 次结果正确率（或差错率）的平均值作为最终结果。本文对各参数分别进行 10 次交叉验证，并选取 10 次差错率的平均值作为对算法精度的估计值。

2.2 融合模拟退火的随机森林算法描述

为使随机森林算法在房价评估问题中达到最佳性能，本文提出将模拟退火与随机森林算法进行融合，提高参数调优的效率，降低算法误差率。

模拟退火融合而成的不同算法，评价函数的选择不同，过于复杂的评价函数会增加算法的消耗，不利于算法的运行，而过于简单的评价算法可能存在准确度不够的情况。就房价评估而言，预测误差是评价函数较优的选择。考虑到评价函数的对比情况，选取预测值和真实值差值与预测值比值的绝对值作为评价函数并取其均值。评价函数表示如下：

$$C(x)=\left|\frac{(y_pred-y_test)}{y_pred}\right| \quad (1)$$

其中：*y_pred* 表示预测值；*y_test* 表示真实值。评价函数值越小，算法预测误差越小；评价函数值越大，算法预测误差越大。

模拟退火算法中，初始温度设置为 T，本文 T 值取其默认值 1 000，将初始值 *x*，初始温度 T，初始评价函数值 $C(x)$ 输入算法内，开始迭代，同时创建新的参数 *x'*，并计算 $C(x')$，比较 $C(x)$ 与 $C(x')$。当误差值减小，即 C(*x'*) 小于 $C(x)$ 时，接受新的参数值 *x'*，或者在波尔兹曼概率内接受新的参数值 *x'*；当所有取值范围内的参数都进行迭代后，降低温度；当经过多次迭代与降温后，若达到最小温度或没有更优解，结束降温，输出参数值。融合模拟退火的随机森林参数调优部分伪代码如算法 1 所示。

算法 1

SA_Parameters 伪代码如下：

输入：初始温度 T,初始参数值 *x*，最小温度 Tmin。

输出：调优参数值 *x*。

```
1. Initizlize   k,m,p   # k 为迭代次数 m 为步长，p 为概率阈值
2. for t=T to Tmin do
3.    for i=1 to k do
4.       x'=x+m
5.       if c(x')<c(x) or exp(-c(x')/T)>p
6.          x=x'
7.       end if
8.    end for
9. end for
10. OutPut x
```

由此得到的 *x* 为调优参数的序列集，融合模拟退火的随机森林算法，替代了传统随机森林的参数选择算法，可以高速有效进行参数调优，有利于提升整个算法运行效率与算法预测准确率。

2.3 融合模拟退火的随机森林房价评估算法构建

将融合模拟退火的随机森林算法应用于房价评估数据，训练出新的房价评估算法。融合模拟退火的随机森林房价评估算法首先将房屋数据集整理为 $D=\{(x_1,y_1),(x_2,y_2),\ldots(x_m,y_m)\}$ 的形式作为输入集进行训练，数据集中每个子集 *x* 代表与房价相关的各项特征，*y* 为此房价真实值；然后确定决策树总数 N，分裂节点数 *k*，确定特征数 *p*，进入算法训练，具体训练过程如下：

输入： $D=\{(x_1,y_1),(x_2,y_2),\ldots(x_m,y_m)\}$ 。

输出：房价评估算法 $f(x)$ 。

a) 确定决策树总数 N，超参数 *k*，特征数 *p*

b) 对每个决策树做如下处理：

for *i*=1 to N

(a) 从 D 中进行 *m* 次有放回随机采样，形成 *m* 个新的采样集 Dt。

(b) 将采样集作为输入训练决策树 Gt，在决策树的分裂节点处从 *p* 个特征中选取 *k* 个，再采取 Gain_σ 计算方差最小值的方法，找到最优特征及阈值作为分类变量，当前节点第 *k* 个特征值小于当前特征划分阈值被划分到左节点，其余被划分为右节点。重复此步骤知道所有节点都已经被训练或被标记为叶子节点。

end for

c) 每棵决策树都进行训练后组成房价评估算法 $f(x)$ 输出。

3 实验分析与结果

3.1 实验数据集

本文选取 kaggle 竞赛中美国某地区成交房屋数据，数据包括 81 个特征变量，房屋属性值包含房屋类型 MSSubClass、小区类型 MSZoning、直线距离 LotFrontage、房屋面积 LotArea、月售额 MoSold、年售额 YrSold、销售类型 SaleType、销售状态 SaleCondition、销售价格 SalePrice 等 81 个房屋属性，涵盖了房屋内部基本属性，与周边环境状况，能够全方位的展示房屋基本信息。房屋部分数据集如表 1 所示。本文将 SalePrice 作为 *y* 变量，其他特征作为 x 变量进行算法训练。由表 1 可得，*x* 变量中包含部分分类数据，本文采用 one-hot 编码对这部分数据进行特征化处理。缺失值采用均值进行填充，在进行数据平滑化处理后，将数据划分为训练集与测试集，训练集占比 0.7。

表 1 实验部分数据集

Table 1 Part of experimental data set

Id	MSSubClass	MSZoning	LotFrontage	LotArea	...
1	60	RL	68	11250	...
2	70	RL	60	9550	...
3	60	RL	84	14260	...
4	20	RL	75	10084	...
5	50	RM	51	6120	...
...	...	...	...	...	...

3.2 参数敏感性测试

随机森林算法敏感性参数测试，各参数取值如表 2 所示。

表 2 参数敏感性实验取值

Table 2 Values of parameter sensitivity test

算法参数	参数取值
n_estimators	100,200,300,500,700,1000,1200
max_features	0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9
min_samples_split	2,3,4,5,6,7,8,9,10
min_samples_leaf	2,3,4,5,6,10
max_leaf_nodes	100,150,200,300,400,500,700,1000,1200
max_depth	2,4,6,8,10,20,60

本文取 10 次 10 折交叉验证差错率的均值作为评价标准，值越小，算法的准确率越高。在保证其他参数为默认参数的前提下，选取单一参数不同取值，形成纵坐标为 10 次 10 折交叉验证差错率的均值（cv_error），横坐标为参数不同取值的图像。观察图像走势，若误差率走势明显或在某个值后误差率基本保持不变，则认为参数不敏感；若误差率在随着参数变动趋势不稳定，在多处均出现最小值，则认为参数敏感。各参数对算法敏感性结果如图 1~6 所示。

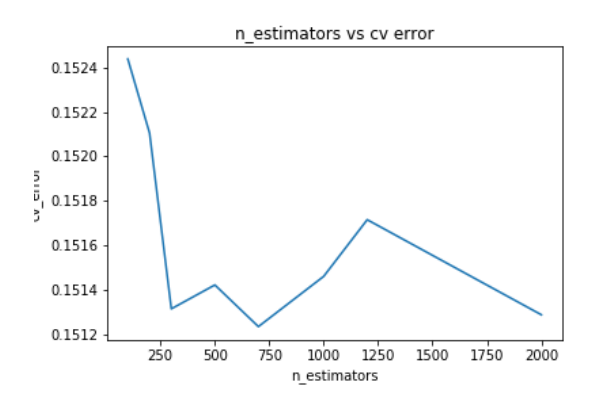


图 1　n_estimators 参数敏感性测试结果

Fig. 1　Sensitive test of n_estimators

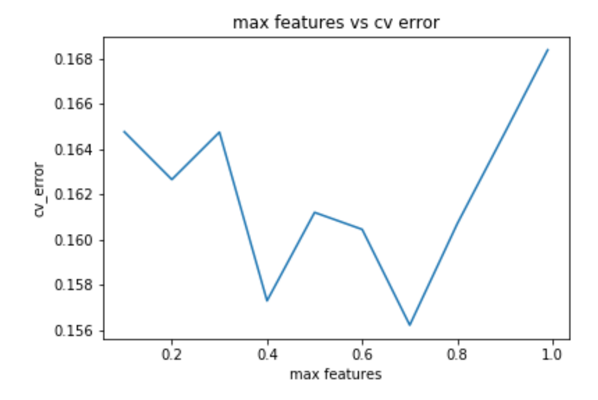


图 2　max_features 参数敏感性测试结果

Fig. 2　Sensitive test of max_features

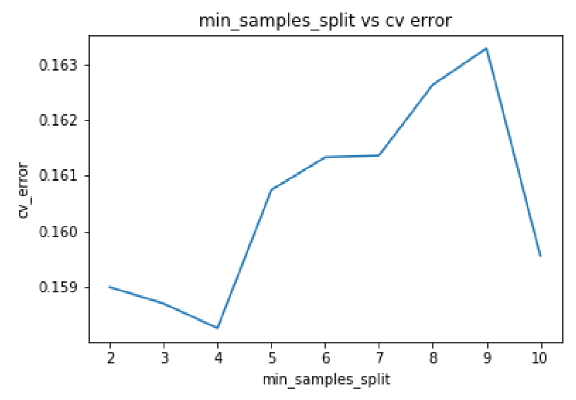


图 3　min_samples_split 参数敏感性测试结果

Fig. 3　Sensitive test of min_samples_split

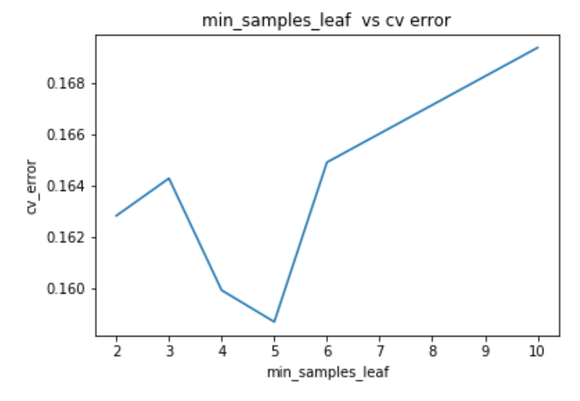


图 4　min_samples_leaf 参数敏感性测试结果

Fig. 4　Sensitive test of min_samples_leaf

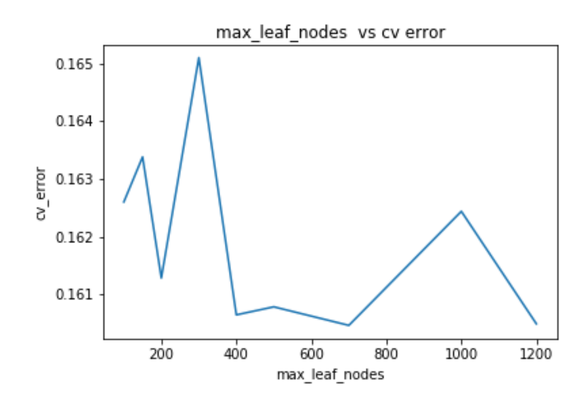


图 5　max_leaf_nodes 参数敏感性测试结果

Fig. 5　Sensitive test of max_leaf_nodes

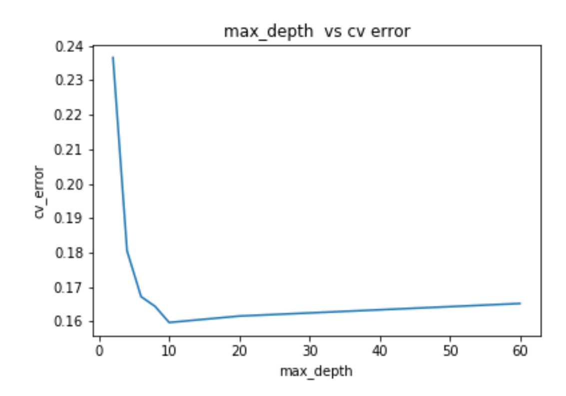


图 6　max_depth 参数敏感性测试结果图

Fig. 6　Sensitive test of max_depth

由图 1~6 可知，经过 10 次 10 折交叉验证的参数敏感性测试，发现 n_estimators、max_feature、max_leaf_nodes 三个参数并不存在唯一最小值使得算法最优，即算法训练数据变化时，三个参数的取值变化对交叉验证差错率影响很大，因此本文认为这三个参数对此回归算法敏感，作为参数优化选项。

min_samples_split 即划分节点所需最小样本数只在四个左右交叉验证差错率最低,min_samples_leaf 即叶子节点最小样本数只在五个左右交叉验证差错率最低，max_depth 即决策树的最大深度在大于 10 后对交叉验证差错率影响相差不大。实验结果表明 min_samples_split、min_samples_leaf、max_depth 三个值对回归算法不敏感，因此不做参数优化，且各参数取值为 min_samples_split=4、min_samples_leaf=5、max_depth=10。

3.3　参数设置

本文根据参数敏感性测试结果，分别采用随机搜索算法、网格搜索算法和模拟退火算法对 n_estimators，max_features，max_leaf_nodes 三个参数进行优化，取值为表 2 中三个参数范围。

这里借助 sklearn 中的随机搜索算法与网格搜索算法包，通过设定随机森林算法与需要调优参数的取值范围进行参数选择。同时，再次应用交叉验证计算算数平均值（记为 mean_validation_score）并与系统运行时间（记为 times）共同作为参考数据，比较各项算法性能。Mean_validation_score 反映了调优后整个算法的预测能力，即通过不同算法选出的参数组合对算法性能的优化情况，系统运行时间 times 反映算法的时效性。三种算法参数选择结果与算法性能比较情况如表 3 所示：

表 3 参数调优结果与算法性能

Table 3 Results of parameter tuning and algorithm performance

算法	参数取值	time/s	mean_validation_score
网格搜索	{1200,500,0.3}	968.02	0.887
随机搜索	{300,300,0.5}	146.512	0.862
模拟退火	{300,500,0.3}	125.443	0.884

由表 3 可得随机搜索算法的运行时间为 125.443 s，模拟退火算法的运行时间为 146.512 s，网格搜索算法的运行时间为 968.02 s，在系统运行时间方面，随机搜索算法与模拟退火算法优于网格搜索算法。从交叉验证平均得分可以看出，网格搜索算法的准确度最高为 0.887，也验证了网格搜索算法高运行时间、高准确度的特点。在运行时间近似的情况下，模拟退火算法参数调优后，算法准确度为 0.884 优于随机搜索算法。综合算法运行时间与算法准确度，模拟退火算法为最优算法，可以弥补传统算法的不足，能够达到快速有效地选择最优参数的作用。

3.4 评价指标

本文选取两类算法考量指标，第一类采用回归算法常见评定指标 MSE、RMSE、R2 三个指标进行评定。其中 MSE（均方误差）代表预测值与真实值的误差平方的期望值，MSE 越小说明算法具有更好的精度；RMSE（均方误差根）是 MSE 的平方根，便于在视图中观察；R^2（拟合优度）反映了自变量对因变量的可解释性，取值小于等于 1，且 R^2 越大越好，如果 R^2 小于 0，则预测算法不如基准算法。三个指标分别定义如下：

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{2}$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \tag{3}$$

$$R^2=1-\left(\sum_{i=1}^{n}(\hat{y}_i-y_i)^2\right)\Big/\left(\sum_{i=1}^{n}(\bar{y}_i-y_i)^2\right) \tag{4}$$

式(2)~(4)中：n 为数据集个数；y_i 为评估结果；$\hat{y}_i$ 为真实值。

另外，本文将算法拟合情况作为第二考量指标，并以散点图的方式进行展示，散点图以真实值为横坐标，预测值为纵坐标，在 $y=x$ 这条线上，预测值与真实值相同，越多的点聚集在 $y=x$ 线上，算法拟合度越好，预测准确度越高。

3.5 算法对比与分析

经过融合模拟退火的随机森林算法，调优参数 n_estimators、max_features、max_leaf_nodes 的最终取值分别为 300、500、0.3。本文首先对融合模拟退火算法前后算法性能进行对比，结果如表 4 所示：

表 4 算法调优前后对比结果

Table 4 Comparison results of algorithm tuning

	参数调优前	参数调优后
MSE	0.039792505	0.02151499
RMSE	0.199480587	0.14667991
R^2	0.762048123	0.87134426

由表 4 可知，参数调优前 MSE 的值为 0.039 792 505，进行参数调优后降到了 0.021 514 99，算法精度得到了提升；RMSE 的值与 MSE 变化相同。拟合度的值从 0.762 048 123 提升到 0.871 344 26，说明自变量对因变量的可解释性增强，即算法拟合情况更优，算法预测准确度增强。为进一步观察算法参数调优前后拟合情况，本文以真实值为横坐标，预测值为纵坐标制作散点图，参数调优前后算法的拟合情况如图 7、8 所示。

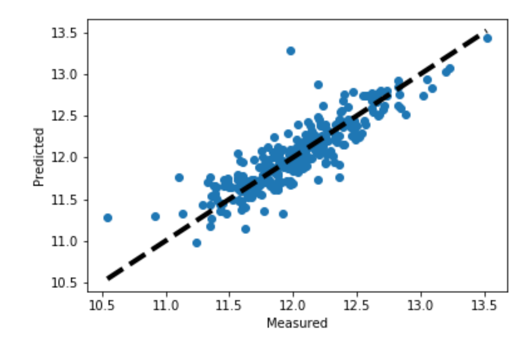


图 7 参数调优前算法拟合情况

Fig. 7 Fitting situation of algorithm before parameter tuning

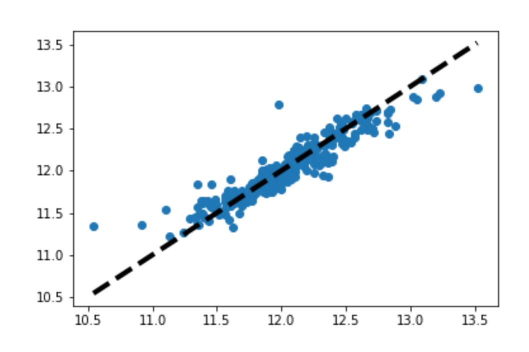


图 8 参数调优后算法拟合情况

Fig. 8 Fitting situation of algorithm after parameter tuning

图 7 中数据点分布较散，且预测值并不稳定。从图 8 中可以看出，进行参数调优后，数据点更加集中在 $y=x$ 附近，对比分析可得，参数调优后算法数据点分布更加集中，算法拟合情况好于参数调优前，这也验证了 MSE、RMSE、R2 三个值的变化，证明模拟退火算法进行参数调优后，算法准确率有所提升。

3.6 与其他算法对比分析

融合模拟退火的随机森林算法与房价评估常用算法 BP 神经网络算法、支持向量机算法形成的房价评估算法在三个量化评价指标上对比的实验结果如表 5 所示。

表 5 各类评估算法对比结果

Table 5 Comparison results of algorithms

	BP	SVM	SA_RF
MSE	0.03204	0.04399	0.02269
RMSE	0.17899	0.20973	0.15062
R^2	0.80842	0.73696	0.86434

融合模拟退火的随机森林算法与房价评估常用算法 BP 神经网络算法、支持向量机算法形成的房价评估算法的拟合情况如图 9~11 所示。

由表 5 及图 9~11 可知，融合模拟退火的随机森林房价评估算法（图中简称为 SA_RF）MSE 值最小为 0.023，BP 神经网络评估算法（MSE=0.032）与支持向量机评估算法（MSE=0.044）均大于此值；新算法的拟合优度值 R2 为 0.864 大于 BP 神经网络评估算法（R2=0.808）与支持向量机评估算法（R2=0.737），拟合情况图中，新模型的数据点更多的聚集在 y=x 这条线上，与 R2 的大小情况相符。综合各项指标值，新算法的误差值最小，拟合程度最高。

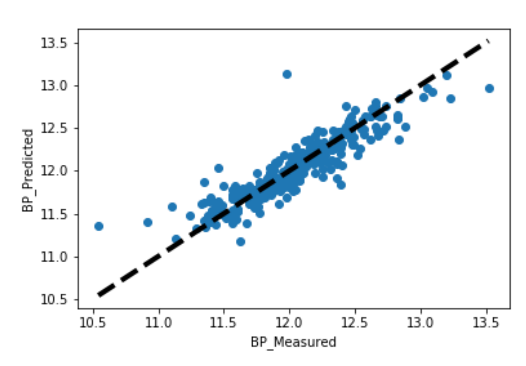


图 9 BP 神经网络评价指标值及算法拟合情况

Fig. 9 Evaluation index of BP neural network and algorithm fitting diagram

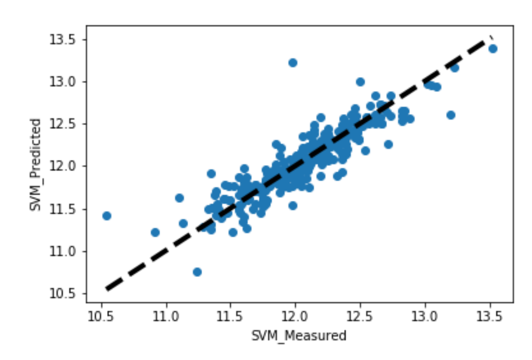


图 10 支持向量机评价指标值及算法拟合情况

Fig. 10 Evaluation index of SVM and algorithm fitting diagram

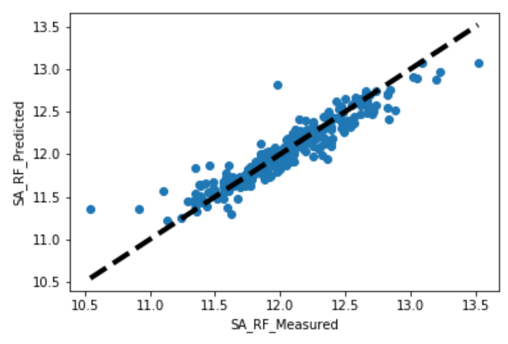


图 11 SA_RF 评价指标值及算法拟合情况图

Fig. 9 Evaluation index of SA_RF and algorithm fitting diagram

4 结束语

通过分析传统房价评估方法，本文提出一种融合模拟退火的随机森林算法进行房价评估，形成房价评估模型。通过对比融合模拟退火算法前后模型以及其他房价评估常用模型，发现融合模拟退火的随机森林房价评估模型拟合情况更优，评估误差率更低。

然而本文仍存在一些缺陷，如模拟退火算法的初始温度 T，只是设置为常用温度 1 000，并未作具体讨论。在接下来的工作中，将对初始温度进行考量，选取出最适合房价评估模型的温度；同时，也会结合我国的国情特色将模型应用到我国房价数据研究。

参考文献：

[1] Hagan M T, Menhaj M B. Training feed forward networks with the Marquardt algorithm [J]. IEEE Trans on Neural Networks, 1994, 5: 989-993.

[2] Gu Jirong, Zhu Mingcang, Jiang Liuguangyan. Housing price forecasting based on genetic algorithm and support vector machine [J]. Expert Systems with Applications, 2011, 38: 3383-3386.

[3] Antipov E A, Povskaya E B. Mass appraisal of residential apartments: an application of random forest for valuation and a CART based diagnostics [J]. Expert Systems with Applications, 2010, 12 (22): 1-18.

[4] Bernard S, Heutte L, Adam S. Influence of hyper parameters on Random Forest accuracy [C]// Proc of the 8th International Workshop on Multiple Classifier System. Berlin: Springer-Verlag, 2009: 171-180.

[5] 温博文，董文瀚，解武杰，等. 基于改进网格搜索算法的随机森林参数优化化 [J]. 计算机工程与应用, 2018, 54 (10): 154-157. (Wen Bowen, Dong Wenhan, Xie Wujie, *et al*. Parameter optimization method for random forest based on improved grid search algorithm [J]. Computer Engineering and Applications, 2018, 54 (10): 154-157.)

[6] Wosniack M E, Raposo E P, Viswanathan G M, *et al*. A parallel algorithm for random searches [J]. Computer Physics Communications, 2015, 196(11): 390–397.

[7] Adnan M N, Islam M Z. Optimizing the number of trees in a accuracy using a genetic algorithm [J]. Knowledge-Based Systems, 2016, 110: 86-97.

[8] 李毓，张春霞. 基于 out—of-bag 样本的随机森林算法的超参数估计 [J]. 系统工程学报，2011, 26 (4): 566-572. (Li Yu. Zhang Chunxia. Estimation of the hyper-parameter in random forest based on out-of-bag sample [J]. Journal of System Engineering, 2011, 26 (4): 566-572.)

[9] Pushpalatha C B, Harrison B P, Sezen S, *et al*. Optimizing event selection with the random grid search [J]. Computer Physics Communications, 2018, 228: 245-257.

[10] 袁梅. 成本法在房屋评估中的应用及分析 [J]. 上海国资，2013, 12: 69-70. (Yuan Mei. Application and analysis of cost method in housing appraisal [J]. Capital Shanghai, 2013, 12: 69-70.)

[11] 李秉坤，孙秀. 市场比较法在土地估价应用中存在的问题及对策建议 [J]. 对外经贸，2015, 10: 120-121. (Li Bingkun, Sun Xiu. The problem and advise of the market comparision methd used in the land valuation [J]. Foreign Economic Relations & Trade, 2015, 10: 120-121.)

[12] 屠蕴雯. 收益法及其在资产评估中的应用 [J]. 科技情报开发与经济, 2009 (3): 131-133. (Tu Yunwen. Discussion on the income approach and its application in assets evaluation [J]. Sci-tech Information Development & Economy, 2009 (3): 131-133.)

[13] 张所地，范新英. 基于面板分位数回归模型的收入、利率对房价的影响关系研究 [J]. 数理统计与管理，2015, 34: 1057-1065. (Zhang Suodi, Fan Xinying. An empirical research about the dynamic influence of income andinterest rate on the housing price based on quantile regression for panel data [J]. Journal of Applied Statistics and Management, 2015, 34: 1057-1065.)

[14] 郭贯成，熊强，汪勋杰. 土地供应政策对房价影响的 GWR 模型分析 [J]. 南京农业大学学报：社会科学版，2014, 14 (5): 91-96. (Guo Guancheng, Xiong Qiang, Wang Xunjie. Effect of land supply policy on China's housing price based on geographically weighted regression model [J]. Journal of Nanjing Agricultural University：Social Sciences Edition , 2014, 14 (5): 91-96.)

[15] Breiman L. Random forests [J]. Mach Learn, 2001, 45: 5-32.

[16] 沈学华，周志华，吴建鑫，等. Boosting 和 Bagging 综述 [J]. 计算机工程与应用，2000, 12: 31-32, 40. (Shen Xuehua, Zhou Zhihua, Wu Jianxin, *et al*. Survey of Boosting and Bagging [J]. Computer

Engineering and Applications, 2000, 12: 31-32, 40.)

[17] 郑延斌，王林林，席鹏雪，等. 基于蚁群算法及博弈论的多 agent 路径规划算法 [J]. 计算机应用，2018 (9): 1-9. (Zheng Yanbin, Wang Linlin, Xi Pengxue, *et al*. Multi agent path planning algorithm based algorithm and game theoryon ant colony [J]. Journal of Computer Applications, 2018 (9): 1-9.)

[18] 巩敦卫，曾现峰，张勇. 基于改进模拟退火算法的机器人全局路径规划 [J]. 系统仿真学报，2013 (3): 480-488. (Gong Dunwei, Zeng Xianfeng, Zhang Yong. Global path planning method of robot based on modified simulated annealing algorithm [J]. Journal of System Simulation, 2013 (3): 480-488.)

[19] 何庆，吴意乐，徐同伟. 改进遗传模拟退火算法在 TSP 优化中的应用 [J]. 控制与决策，2018 (2): 219-225. (He Qing, Wu Yile, Xu Tongwei. Application of improved genetic simulated annealing algorithm in TSP optimization [J]. Control and Decision, 2018 (2): 219-225.)

[20] 傅文渊，凌朝东. 布朗运动模拟退火算法 [J]. 计算机学报, 2014 (6): 1301-1308. (Fu Wenyuan, Ling Chaodong. Brownian motion based simulated annealing algorithm [J]. Chinese Journal of Computers, 2014 (6): 1301-1308.)

[21] Oshiro T M, Perez P S, Baranauskas J A. How many trees in a random forest [C]// Lecture Notes in Computer Science: International Workshop on Machine Learning and Data Mining in Pattern Recognition, 2012, 7376: 154-168.

[22] Rodriguez-Galiano V F, Ghimire B, Rogan J, Chica-Olmo M, Rigol-Sanchez J P. An assessment of the effectiveness of a random forest classifier for landcover classification [J]. ISPRS Journal of Photogrammetry and Remote Senssing, 2012, 67 (1): 93-104.

[23] Díaz-Uriarte R, De Andres S A. Gene selection and classification of microarray data using random forest [J]. BMC Bioinformatics, 2006, 7: 1-3.